# Dynamic Programming for Epistemic Uncertainty in Markov Decision Processes

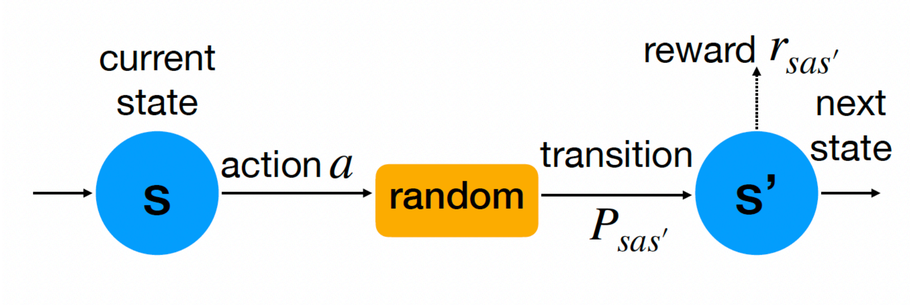
**Axel Benyamine**, Julien Grand-Clément, Marek Petrik, Michael I. Jordan, Alain Durmus

# Outline

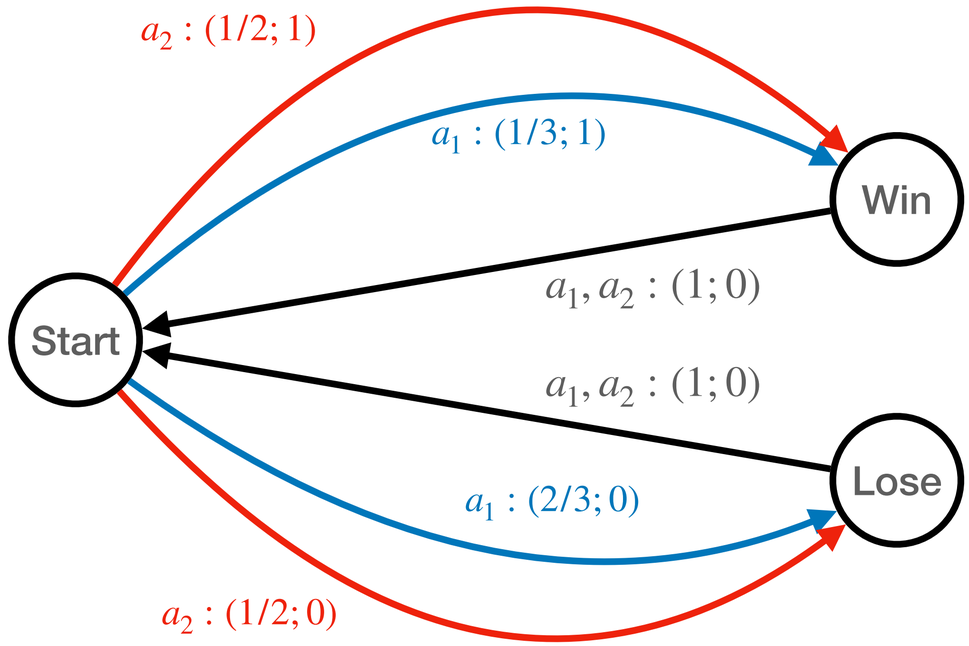
- MDPs and Motivation
- Dynamic Programming
- Risk-Measures
- Ambiguity-Averse MDPs
- DP-compatible risk-measures
- Main Results
- Discussion and Conclusion

# MDPs and Motivation

# Markov Decision Processes




MDP with three states and two actions. The edges $(s, a, s')$ are labeled with pairs $(P(s, a, s'), r(s, a, s'))$.




- Finite set of states and actions: $\mathcal{S}, \mathcal{A}$
- Reward $r(s, a, s') \in \mathbb{R}$, discount factor $\gamma \in [0,1)$
- Transition probabilities $P = (P(s, a, s'))$
- History-dependent policy $\pi \in \Pi_{\mathsf{H}}$: maps all finite histories to $\Delta(\mathcal{A})$
- Stationary policy $\pi \in \Pi_{\mathsf{S}}$: maps $\mathcal{S}$ to $\Delta(\mathcal{A})$
- Objective: Evaluate and Optimize $V^{\pi,P}(s) = \mathbb{E}^{\pi,P}_{\tilde{S}_0=s}\left(\sum_{t=0}^{\infty} \gamma^t r(\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1})\right)$

# Problem: Epistemic Uncertainty in MDPs

**Nominal MDPs:**
transition kernel $P$ is known and fixed

**In practice:** $P$ is estimated $\Rightarrow$ model errors

**High-stakes settings:**

- Healthcare (scheduling, treatment)
- Vehicle routing
- Finance / Energy

**Existing litteratures:**

- **Robust MDPs:** studies $\inf_{P \in \mathcal{P}} V^{\pi,P}(s)$
- **Optimistic MDPs:** studies $\sup_{P \in \mathcal{P}} V^{\pi,P}(s)$
- **Multi-model MDPs:** studies $\text{avg}_{P \in \mathcal{P}} V^{\pi,P}(s)$
- **Percentile Opt.:** VaR (∼quantile)

**Each introduced independently, with *ad hoc* analysis.**

**MAIN RESEARCH QUESTION**: Provide a **unified picture** of what is possible.

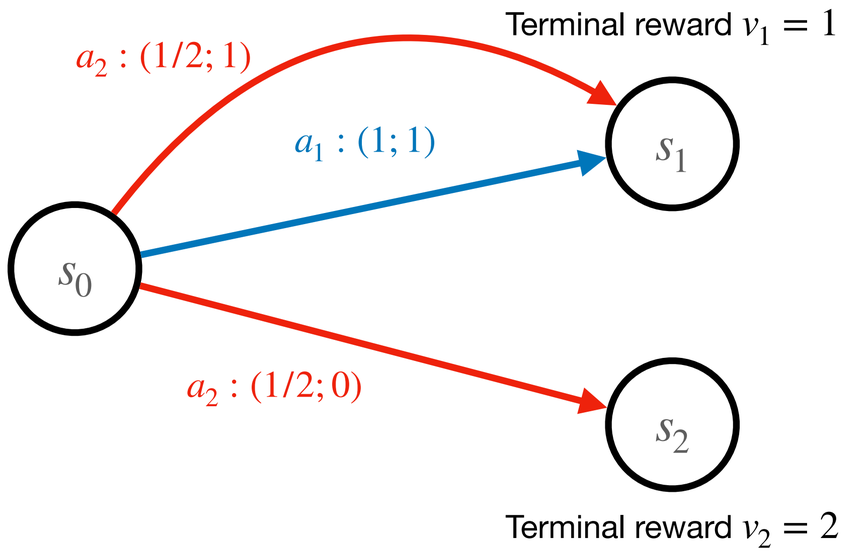
# Dynamic Programming

# Bellman Operator

- Policy $(\lambda_1, \lambda_2)$: takes $a_i$ with proba $\lambda_i$
- **Expected reward after one time step:**

$$\lambda_1(1+\gamma) + \lambda_2\left(1/2(1+\gamma) + 1/2(0+2\gamma)\right)$$

MDP with three states and two actions. The edges $(s, a, s')$ are labeled with pairs $(P(s, a, s'), r(s, a, s'))$.




**Bellman operator:**

Input: Reward $v$ for playing $[t+1, \ldots, \infty)$ rounds

Output: Reward $T^{\pi,P}v$ for playing $[t, \ldots, \infty)$ rounds

$$T^{\pi,P}v(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} P(s,a,s')\Big(r(s,a,s') + \gamma v(s')\Big)$$

# Dynamic Programming for Nominal MDPs

**Value function**:

$$V^{\pi,P}(s) = \mathbb{E}_s^{\pi,P}\left(\sum_{t=0}^{\infty} \gamma^t r(\tilde{S}_t, \tilde{A}_t, \tilde{S}_{t+1})\right)$$

**Bellman operator:**

$$T^{\pi,P}v(s) = \sum_{a\in\mathcal{A}} \pi(s,a) \sum_{s'\in\mathcal{S}} P(s,a,s')\big(r(s,a,s') + \gamma v(s')\big)$$

**Optimal Value function**: $V^P(s) = \sup_{\pi\in\Pi_H} V^{\pi,P}(s)$

**Optimal Bellman operator:** $T^P v(s) = \sup_{\pi\in\Pi_S} T^{\pi,P}v(s)$

**Dynamic Programming Principle [Puterman, 2014]:**

For any MDP and $\pi \in \Pi_S$:

$$V^{\pi,P} = T^{\pi,P}V^{\pi,P} \quad \text{and} \quad V^P = T^P V^P$$

$\Rightarrow$ Stationary optimal policy · Value / Policy iteration · Exponential Convergence

**Question:** Does a similar principle hold when $P$ is uncertain ?

# Risk Measures

# Risk Measures

**Definition:**
A risk measure $\rho$ maps a random variable to a scalar.

**Key property:** Law-invariance

$$\mathscr{L}(\tilde{X}) = \mathscr{L}(\tilde{Y}) \implies \rho(\tilde{X}) = \rho(\tilde{Y})$$

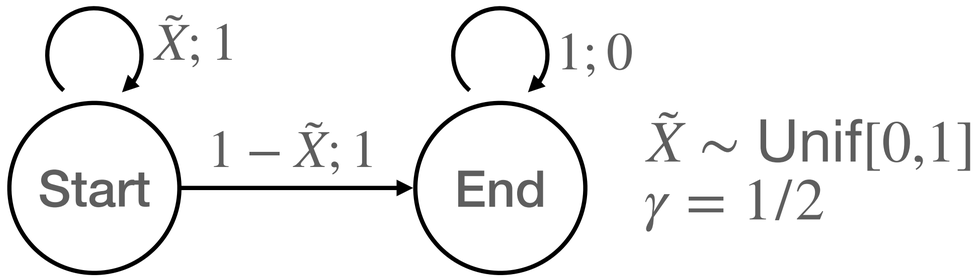
**Standard Examples:**

- Expectation $\mathbb{E}[\tilde{X}]$
- Essential Infimum $\operatorname{ess\,inf}[\tilde{X}] = \inf \operatorname{supp}(\tilde{X})$
- Essential Supremum $\operatorname{ess\,sup}[\tilde{X}] = \sup \operatorname{supp}(\tilde{X})$
- $\mathrm{VaR}_\alpha(\tilde{X})$, $\mathrm{CVaR}_\alpha(\tilde{X})$, $\mathrm{ERM}_\beta(\tilde{X})$

Most standard examples are **law-invariant** and **monetary** (monotone + translation-invariant)

# Ambiguity-Averse MDPs

# Framework Definition




**Definition**:
An **Ambiguity-Averse MDP** is a tuple $(\mathcal{A}, \mathcal{S}, r, \nu, \gamma)$ with $\nu$ a distribution over transition kernels.

**Core idea**: treat $\tilde{P}$ as a random variable, the value function $V^{\pi,\tilde{P}}$ is now random and evaluated with a risk-measure $\rho$.

**Assumptions on $\nu$:**

1. $\nu$ has **product structure**: the random variables $\left\{\tilde{P}(s, \cdot, \cdot)\right\}_{s \in \mathcal{S}}$ are mutually independent under $\nu$
2. $\text{supp}(\nu)$ is **convex**

**Two uncertainty models:**

- **Static:** $\tilde{P}_t = \tilde{P}_0$ for all t
  one initial draw, fixed forever
- **Resampled:** $\tilde{P}_t \overset{\text{iid}}{\sim} \nu$
  new draw at every period

# DP-compatible risk-measures

# DP-compatible risk-measures

**A-A Value function:** $V^{\pi,\nu,\rho}(s) = \rho(V^{\pi,\tilde{P}}(s))$

**A-A Bellman operator:** $T^{\pi,\nu,\rho}v(s) = \rho(T^{\pi,\tilde{P}}v(s))$

**A-A Optimal Value function:**

$$V^{\nu,\rho}(s) = \sup_{\pi\in\Pi_H} V^{\pi,\nu,\rho}(s)$$

**A-A Optimal Bellman operator:**

$$T^{\nu,\rho}v(s) = \sup_{\pi\in\Pi_S} T^{\pi,\nu,\rho}v(s)$$

**Question:**

For which $\rho$ do we have

- **Bellman Evaluation Equation (BE):**

  $V^{\pi,\nu,\rho} = T^{\pi,\nu,\rho}V^{\pi,\nu,\rho}$ for any ambiguity-averse MDP and $\pi \in \Pi_S$

- **Bellman Optimality Equation (BO):**

  $V^{,\nu,\rho} = T^{\nu,\rho}V^{\nu,\rho}$ for any ambiguity-averse MDP

# Example : Robust MDPs ($\rho = \operatorname{ess\,inf}$)

**Reformulation of Robust MDPs:**

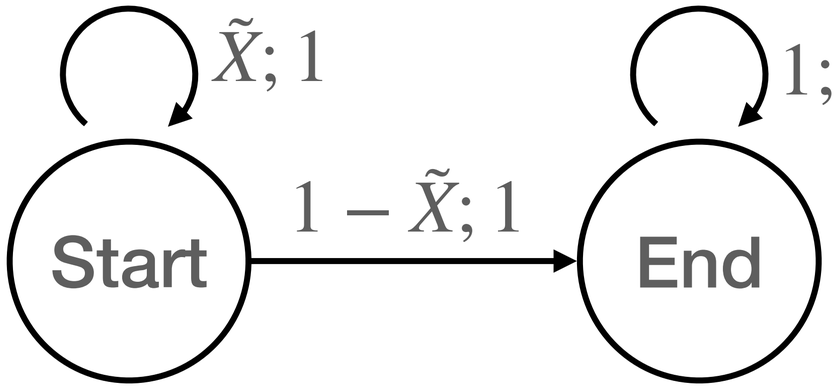
$$\operatorname{ess\,inf}_{\tilde{P}\in\nu} V^{\pi,\tilde{P}}(s) = \inf_{P\in\mathrm{supp}(\nu)} V^{\pi,P}(s)$$

$$\sup_{\pi\in\Pi_H} \operatorname{ess\,inf}_{\tilde{P}\in\nu} V^{\pi,\tilde{P}}(s) = \sup_{\pi\in\Pi_H} \inf_{P\in\mathrm{supp}(\nu)} V^{\pi,P}(s)$$

**Conclusion from RMDP literature [Wiesemann et al., 2013]:**
$\rho = \operatorname{ess\,inf}$ is DP-compatible (i.e. it satisfies both Bellman Equations)

$\Rightarrow$ Stationary optimal policy · Value / Policy iteration · Exponential Convergence

# Counter-Example: Multi-model MDPs for static kernels




$\tilde{X} \sim$ **Unif**$[0,1]$ **static**

one initial draw, fixed forever

$\gamma = 1/2$, **Single action**

$$V^{\pi,\nu,\mathbb{E}^\nu} = \begin{pmatrix} \mathbb{E}\left(\sum_{t\geqslant 0}(1/2)^t \tilde{X}^t\right) \\ \mathbb{E}(0) \end{pmatrix} = \begin{pmatrix} 2\log(2) \\ 0 \end{pmatrix}$$

$$T^{\pi,\nu,\mathbb{E}^\nu} V^{\pi,\nu,\mathbb{E}^\nu} = \begin{pmatrix} 1 + \log(2)/2 \\ 0 \end{pmatrix} \neq V^{\pi,\nu,\mathbb{E}^\nu}$$

**Conclusion:**

$\mathbb{E}$ is **not** DP-compatible for static kernels
(i.e. it doesn't satisfy the Bellman Equations)

Main issue: $\mathbb{E}[\tilde{X}^2] \neq \mathbb{E}[\tilde{X}]^2$ for $\tilde{X} \sim$ Unif$[0,1]$

# Main Results

# Complete Characterization

**Theorem:**

**Static Kernels:** The following are equivalent

(i) $\rho$ satisfies **(BE)** equation (ii) $\rho$ satisfies **(BO)** equation (iii) $\rho \in \{\text{ess inf}, \text{ess sup}\}$

**Resampled Kernels:** The following are equivalent

(i) $\rho$ satisfies **(BE)** equation (ii) $\rho$ satisfies **(BO)** equation (iii) $\rho \in \{\text{ess inf}, \text{ess sup}, \mathbb{E}\}$

**Proof idea:** Exhibits a MDP structure where

- transition to **one** next state $\Rightarrow$ positive homogeneity i.e. $\rho(\alpha\tilde{X}) = \alpha\rho(\tilde{X})$ for $\alpha \geq 0$
- transitions to **two** next states $\Rightarrow$ additivity (independent) i.e. $\rho(\tilde{X} + \tilde{Y}) = \rho(\tilde{X}) + \rho(\tilde{Y})$ for $\tilde{X} \perp\!\!\!\perp \tilde{Y}$

**Surprising (BO) $\Longleftrightarrow$(BE):**

Evaluation over all policies is equivalent to evaluation over the single optimal policy

# Discussion and Conclusion

# Takeaways

**Advantages of DP**

- Stationary optimal policies exist
- Value and Policy iteration at exponential rate

**× Impossibility**

- VaR, CVaR, ERM, EVaR all DP-incompatible
- **Explains why no new tractable model emerged after RMDPs**
- Static ≫ Resampled in difficulty

**Complete Characterization**

Among law-invariant monotone risk-measures, only **robust** ($\operatorname{ess\,inf}$), **optimistic** ($\operatorname{ess\,sup}$) and **neutral** ($\mathbb{E}$, resampled) admit DP.

**Beyond DP**

- Augmented state space
- Nested risk-measures
- Direct policy gradient

# Conclusion

1. **Unifying framework:** ambiguity-averse MDPs recover nominal, robust, optimistic, multi-model and percentile MDPs via the choice of $\rho$.
2. **DP Machinery:** Bellman operators remain contractions, value/policy iteration converge exponentially.
3. **Complete Characterization:** Only $\operatorname{ess\,inf}$, $\operatorname{ess\,sup}$ and $\mathbb{E}$ (resampled) are DP-compatible among law-invariant monotone risk-measures.

**Thank you!**
**arXiv:2602.03381**

# Appendix

# Robust MDPs and $s$-rectangularity

A **Robust MDP** is a tuple $(\mathcal{A}, \mathcal{S}, r, \mathscr{P}, \gamma, \mu)$ with $\mathscr{P} \subseteq \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}}$ a set of possible probability kernels

**Definition: $s$-rectangularity [Wiesemann et al., 2013]**

$\mathscr{P}$ is $s$-rectangular if it decomposes state by state: $\mathscr{P} = \times_{s\in\mathcal{S}} \mathscr{P}_s\,, \qquad \mathscr{P}_s \subseteq \Delta(\mathcal{S})^{\mathcal{A}}$

Examples for 2 states $(s_1, s_2)$ and 1 action $(a)$:

- $\mathscr{P} = \left\{ \left( \begin{pmatrix} P(s_1,a,s_1) : 1/2-\varepsilon \\ P(s_1,a,s_2) : 1/2+\varepsilon \end{pmatrix}, \begin{pmatrix} P(s_2,a,s_1) : \xi \\ P(s_2,a,s_2) : 1-\xi \end{pmatrix} \right) \quad \varepsilon \in [-1/4,1/4], \xi \in [0,1/3] \right\}$ is $s$-rectangular
- $\mathscr{P} = \left\{ \left( \begin{pmatrix} P(s_1,a,s_1) : x \\ P(s_1,a,s_2) : 1-x \end{pmatrix}, \begin{pmatrix} P(s_2,a,s_1) : y \\ P(s_2,a,s_2) : 1-y \end{pmatrix} \right) \quad x+y=1, \quad x,y \in [0,1] \right\}$ is not $s$-rectangular

$\Rightarrow$ In practice: each $P(s,\cdot,\cdot)$ must be estimated independently

# Dynamic Programming for Robust MDPs

**Robust Value function:** $V^{\pi,\mathscr{P},\inf}(s) = \inf_{P\in\mathscr{P}} V^{\pi,P}(s)$

**Robust Optimal Value function:** $V^{\mathscr{P},\inf}(s) = \sup_{\pi\in\Pi_H} V^{\pi,\mathscr{P},\inf}(s)$

**Robust Bellman operator:** $T^{\pi,\mathscr{P},\inf}v(s) = \inf_{P\in\mathscr{P}} T^{\pi,P}v(s)$

**Robust Optimal Bellman operator:** $T^{\mathscr{P},\inf}v(s) = \sup_{\pi\in\Pi_S} T^{\pi,\mathscr{P},\inf}v(s)$

**Dynamic Programming for RMDPs [Wiesemann et al., 2013]:**

If $\mathscr{P}$ is $s$-**rectangular** and **convex**, then for any RMDP and $\pi \in \Pi_S$:

$$V^{\pi,\mathscr{P},\inf} = T^{\pi,\mathscr{P},\inf}V^{\pi,\mathscr{P},\inf} \quad \text{and} \quad V^{\mathscr{P},\inf} = T^{\mathscr{P},\inf}V^{\mathscr{P},\inf}$$

$\Rightarrow$ Stationary optimal policy · Value / Policy iteration · Exponential Convergence

**Question:** Does a similar principle hold if we replace $\inf$ by something else ?

# DP Machinery

**Proposition [Crandall and Tartar, 1980]:**
A monotone translation invariant risk-measure is non-expansive: $|\rho(\tilde{X}) - \rho(\tilde{Y})| \leqslant \operatorname{ess\,sup} |\tilde{X} - \tilde{Y}|$

**Proposition:** If $\rho$ is monotone and translation-invariant, then

- both $T^{\pi,\nu,\rho}$ and $T^{\nu,\rho}$ are **monotone**, **translation-invariant** and **contracting** (i.e. $|T(\tilde{X}) - T(\tilde{Y})| < \gamma \operatorname{ess\,sup} |\tilde{X} - \tilde{Y}|$)
- **Attainability**: For any $v \in \mathbb{R}^{\mathcal{S}}$, there exists $\pi \in \Pi_S$ such that $T^{\nu,\rho} v = T^{\pi,\nu,\rho} v$.

# Summary Table of main variants

| Name | Objective | Risk measure $\rho(\tilde{X})$ | Static Kernels | | | Resampled Kernels | | |
|---|---|---|---|---|---|---|---|---|
| | | | DP | $\pi^\star \in \Pi_S$ | Trac. | DP | $\pi^\star \in \Pi_S$ | Trac. |
| Nominal MDP | $\mu^\top V^{\pi,P}$ | $\mathbb{E}^\nu(\tilde{X})$ with $\nu = \delta_P$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Robust MDP | $\inf_{P\in\mathcal{P}} \mu^\top V^{\pi,P}$ | $\operatorname{ess\,inf}(\tilde{X})$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Optimistic MDP | $\sup_{P\in\mathcal{P}} \mu^\top V^{\pi,P}$ | $\operatorname{ess\,sup}(\tilde{X})$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Multi-model MDP | $\mathbb{E}^\nu(\mu^\top V^{\pi,\tilde{P}})$ | $\mathbb{E}^\nu(\tilde{X})$ | × | × | × | ✓ | ✓ | ✓ |
| Percentile optimization | $\mathsf{VaR}_\alpha(\mu^\top V^{\pi,\tilde{P}})$ | $\mathsf{VaR}_\alpha(\tilde{X})$ | × | × | × | × | × | × |

- **DP**: Bellman Equations hold **Tract.**: polynomial-time solvable
- Static kernels are strictly harder than resampled kernels (more correlation)

# Necessary Conditions to be DP-compatible

**Proposition:** Let $\rho$ be law-invariant and non-constant. If $\rho$ is DP-compatible, then it satisfies

- Identity on constants: $\rho(c) = c$ for all $c \in \mathbb{R}$
- Positive homogeneity: $\rho(\alpha\tilde{X}) = \alpha\rho(\tilde{X})$ for $\alpha \geq 0$
- Additivity (independent): $\rho(\tilde{X} + \tilde{Y}) = \rho(\tilde{X}) + \rho(\tilde{Y})$ for $\tilde{X} \perp\!\!\!\perp \tilde{Y}$

**Proof idea:** Exhibits a MDP structure where

- transition to **one** next state ⇒ positive homogeneity i.e. $\rho(\alpha\tilde{X}) = \alpha\rho(\tilde{X})$ for $\alpha \geq 0$
- transitions to **two** next states ⇒ additivity (independent) i.e. $\rho(\tilde{X} + \tilde{Y}) = \rho(\tilde{X}) + \rho(\tilde{Y})$ for $\tilde{X} \perp\!\!\!\perp \tilde{Y}$

**Already rules out:**

- $\text{ERM}_\beta$ (not positive homogeneous)
- $\text{CVaR}_\alpha$, $\text{VaR}_\alpha$ (not additive independent)